Tutorial: Building InfiniBand clusters with Open Fabrics Software Stack

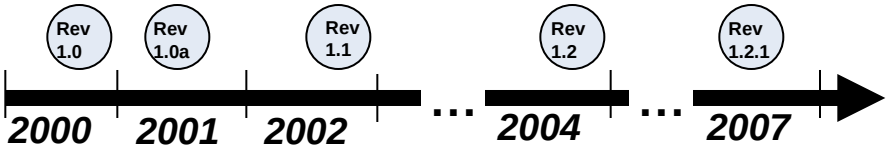
HPC Advisory Council
Lugano, Switzerland Workshop
March 13-15th, 2012

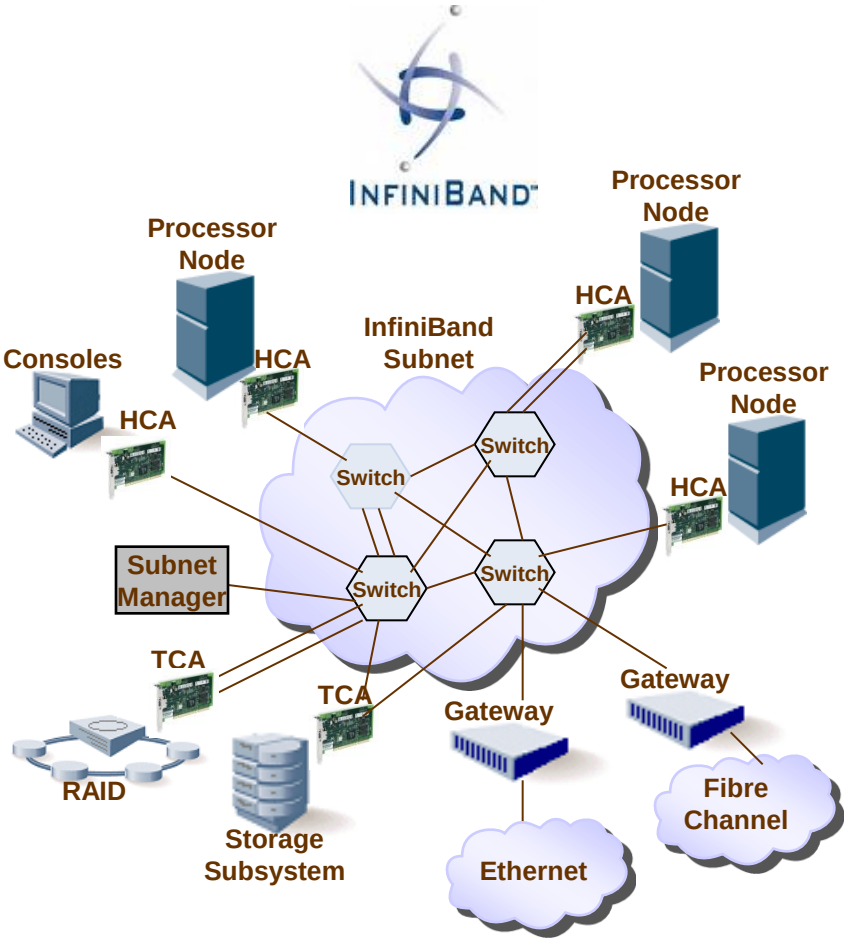
Todd Wilde - Director of Technical Computing and HPC

InfiniBand Overview

The InfiniBand Architecture

- Industry standard defined by the InfiniBand Trade Association
- Defines System Area Network architecture
  - Comprehensive specification:
    from physical to applications





- Architecture supports
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload

3

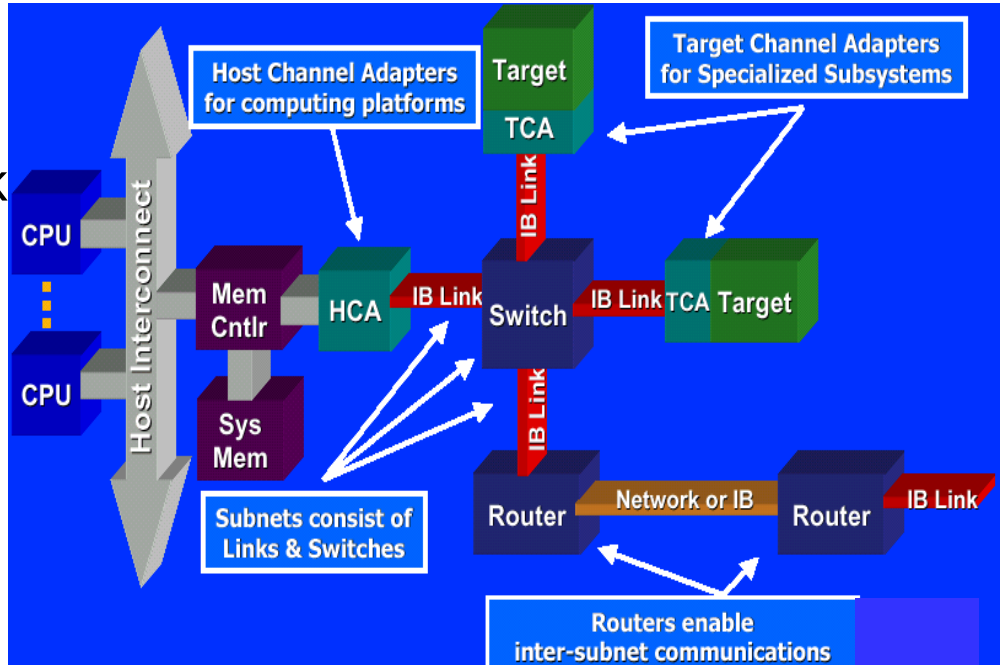
InfiniBand Feature Highlights

- Serial High Bandwidth Links
  - 56Gb/s HCA and switch links
  - Over 12-GB/s data throughput
- Ultra low latency
  - Under 1 us application to application
- Reliable, lossless, self-managing fabric
  - Link level flow control
  - Congestion control to prevent HOL blocking
- Full CPU Offload
  - Hardware Based Transport Protocol
  - Reliable Transport
  - Kernel Bypass (User level applications get direct access to hardware)
- Memory exposed to remote node access
  - RDMA-read and RDMA-write
- Quality Of Service
  - Independent I/O channels at the adapter level
  - Virtual Lanes at the link level
- Cluster Scalability/flexibility
  - Up to 48K nodes in subnet, up to 2^{128} in network
  - Parallel routes between end nodes
  - Multiple cluster topologies possible
- Simplified Cluster Management
  - Centralized route manager
  - In-band diagnostics and upgrades

InfiniBand Hardware

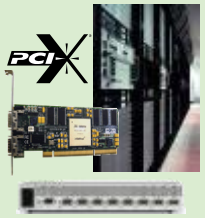
InfiniBand Hardware Components

- Host Channel Adapter (HCA)
  - Device that terminates an IB link and executes transport-level functions and support the verbs interface
- Switch
  - A device that routes packets from one link to another of the same IB Subnet
- Cables
  - QSFP copper or fiber cables used to connect the HCA and switch ports together

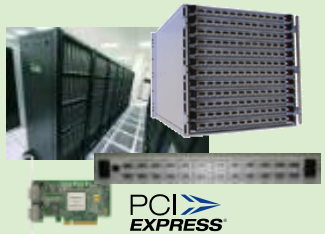


Introducing FDR InfiniBand 56Gb/s Solutions

2002
10Gb/s



2005
20Gb/s



2008
40Gb/s

2011
56Gb/s

Highest Performance, Reliability, Scalability, Efficiency

FDR InfiniBand New Features and Capabilities

Performance / Scalability

- >12GB/s bandwidth, <0.7usec latency
- PCI Express 3.0
- InfiniBand router and IB-Eth/FC bridges

Reliability / Efficiency

- Link bit encoding – 64/66
- Forward Error Correction
- Lower power consumption

FDR 56Gb/s InfiniBand Solutions Portfolio

Modular Switch

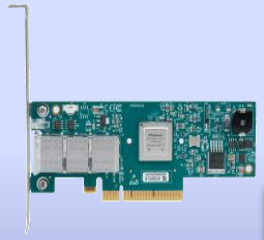
648p

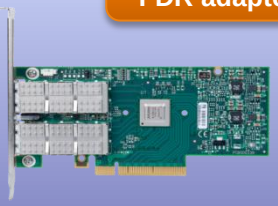
324p

216p

SX6500
Up to 648 ports FDR

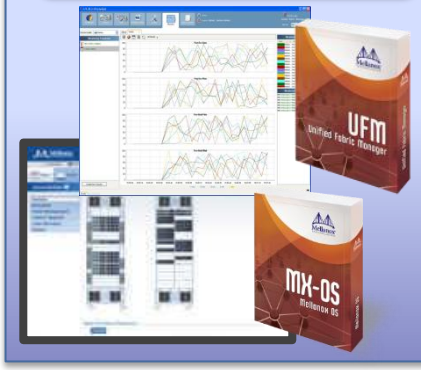
Adapters





Single and dual port FDR adapter card

Management



Mellanox M-1 E-2-E Cluster Support Services

Edge Switch

SX60XX
1U 36 port

SX6036 – 36 ports managed

SX6025 – 36 ports externally managed

Cables

Feature License Keys

UFM Diagnostics

Virtual Protocol Interconnect

Note: Please check availability with your Mellanox representative

Cable Interconnect

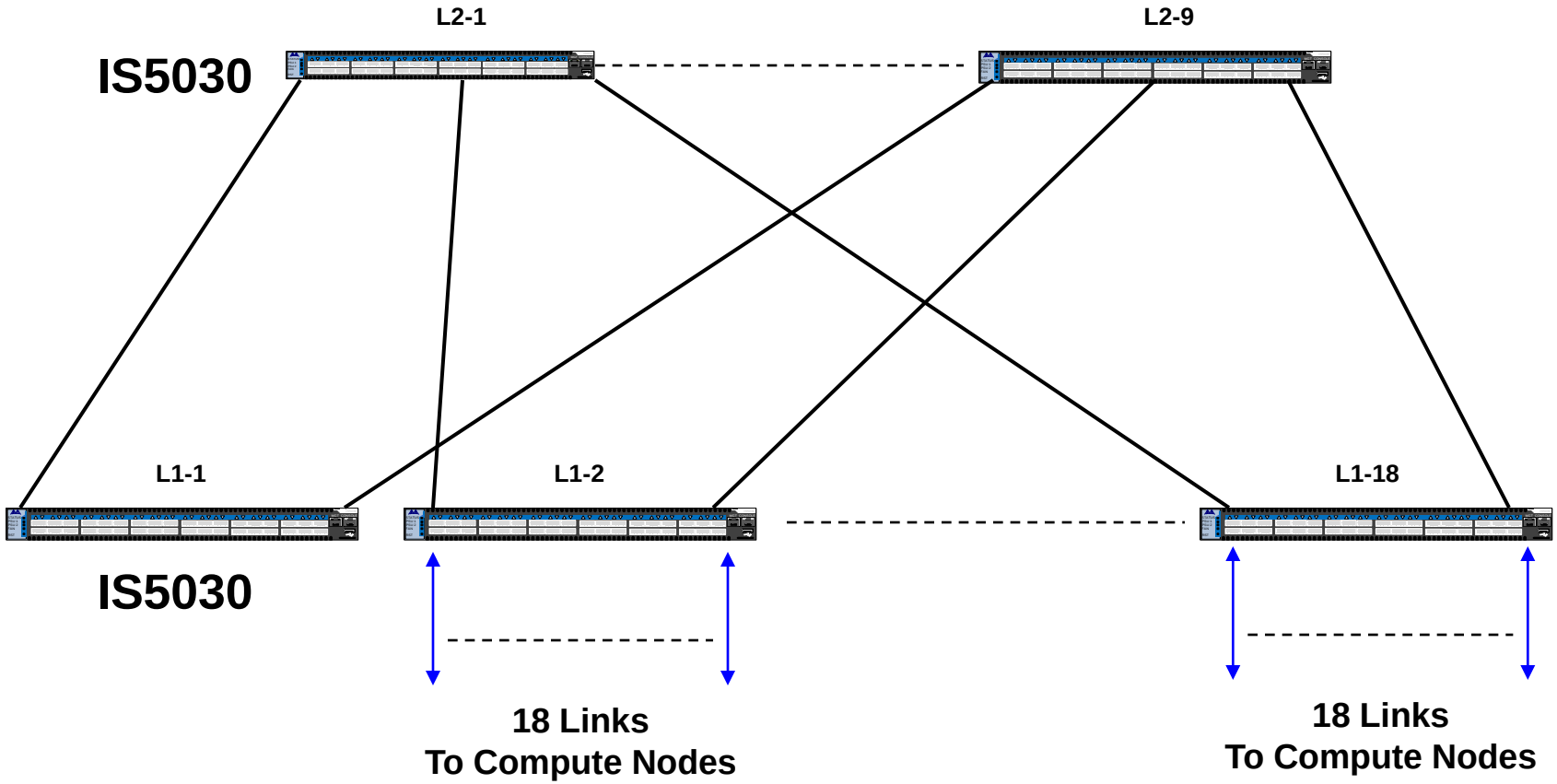
Cluster Topologies

- Topologies used in large clusters
  - Fat-Tree
  - 3D Torus
  - Mesh/Hypercube

- Fat-tree characteristics:
  - Use same BW for all links (or close BW)
  - Many times use same number of ports for all switches
  - Several blocking and full non-blocking configurations possible

- Main issues with fabric design
  - Is the SM capable of routing the fabric?
  - Does it generate credit loops?
  - Are the paths evenly distributed?

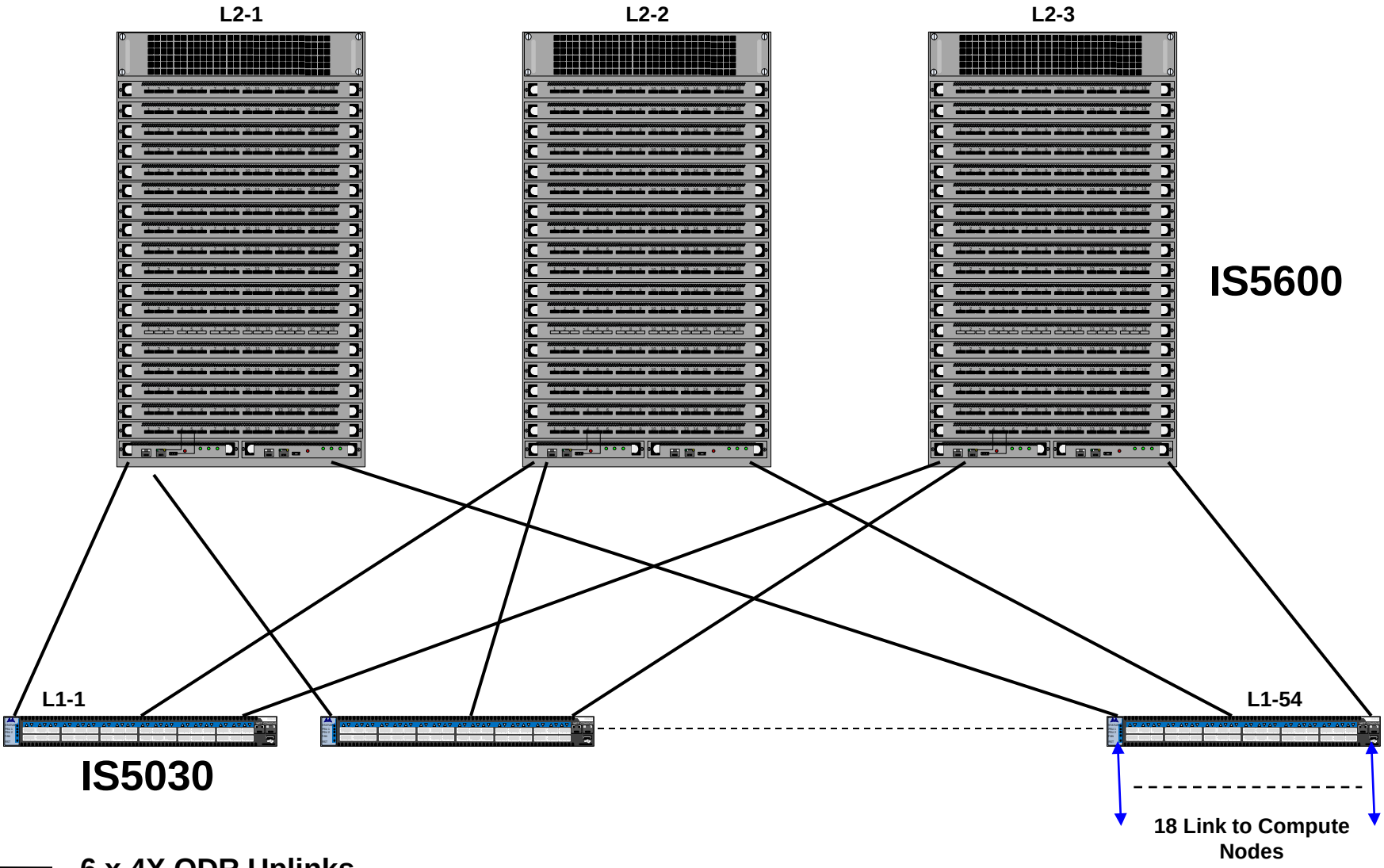
324 Node Full CBB Using 36-port Switches



2 x 4X QDR Uplinks

1 x 4X QDR Uplinks

1944 Node Full CBB using 36/648 Port Switch



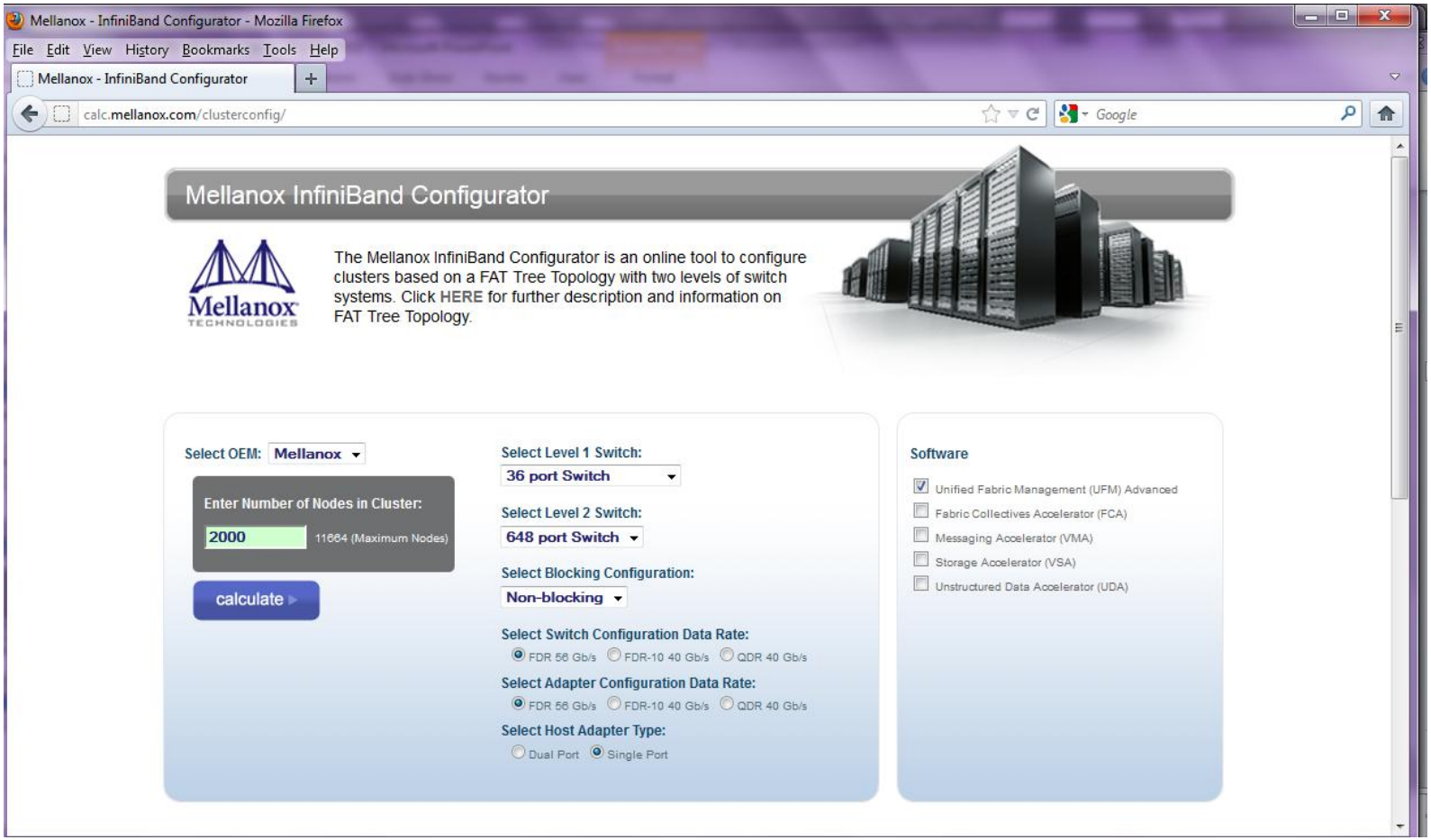
Cabling Tips

- Label all cables between the leaf and the core switches
- Failure to do so will hamper best efforts to isolate, identify and troubleshoot any bad cables when the cluster is brought up.
- Sample cable labeling:

Connection	Near End	Far End
Server-L1 switch	*Node Name* *Leaf# / Slot# / Port#*	*Leaf# /Slot# / Port#* *Node Name*
Switch-Switch	*Leaf# / Slot# / Port#* *Spine# / Slot# / Port#*	*Spine# / Slot# / Port#* *Leaf# / Slot# / Port#*

Cluster Configurator

- Cluster configuration tool available from Mellanox
  - http://calc.mellanox.com/clusterconfig/


Mellanox - InfiniBand Configurator - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Mellanox - InfiniBand Configurator
calc.mellanox.com/clusterconfig/
Google
Mellanox InfiniBand Configurator
The Mellanox InfiniBand Configurator is an online tool to configure clusters based on a FAT Tree Topology with two levels of switch systems. Click HERE for further description and information on FAT Tree Topology.
Select OEM: Mellanox
Enter Number of Nodes in Cluster:
2000
11664 (Maximum Nodes)
calculate
Select Level 1 Switch:
36 port Switch
Select Level 2 Switch:
648 port Switch
Select Blocking Configuration:
Non-blocking
Select Switch Configuration Data Rate:
FDR 56 Gb/s FDR-10 40 Gb/s QDR 40 Gb/s
Select Adapter Configuration Data Rate:
FDR 56 Gb/s FDR-10 40 Gb/s QDR 40 Gb/s
Select Host Adapter Type:
Dual Port Single Port
Software
Unified Fabric Management (UFM) Advanced
Fabric Collectives Accelerator (FCA)
Messaging Accelerator (VMA)
Storage Accelerator (VSA)
Unstructured Data Accelerator (UDA)

Basics of the OFED InfiniBand Stack

MLNX_OFED Installation

- Pre-built RPM install.
  - 1. mount -o rw,loop MLNX_OFED_LINUX-1.4-rhel5.3.iso /mnt
  - 2. cd /mnt
  - 3. ./mlnxofedinstall
- Building RPMs for un-supported kernels.
  - 1. mount -o rw,loop MLNX_OFED_LINUX-1.4-rhel5.3.iso /mnt
  - 2. cd /mnt/src
  - 3. cp OFED-1.4.tgz /root (this is the original OFED distribution tarball)
  - 4. tar zxvf OFED-1.4.tgz
  - 5. cd OFED-1.4
  - 6. copy ofed.conf to OFED-1.4 directory
  - 7. ./install.pl -c ofed.conf

MLNX_OFED Configuration

- Loading and Unloading the IB stack
  - /etc/infiniband/openib.conf controls boot time configuration and other options

```
# Start HCA driver upon boot
ONBOOT=yes

# Load IPoIB
IPOIB_LOAD=yes
```

  - Optionally manually start and stop services
    - /etc/init.d/openibd start|stop|restart|status

IPoIB in a Nut Shell

- Encapsulation of IP packets over IB
- Uses IB as “layer two” for IP
  - Supports both UD service (up to 2KB MTU) and RC service (connected mode, up to 64KB MTU)
- IPv4, IPv6, ARP and DHCP support
- Multicast support
- VLANs support
- Benefits:
  - Transparent to legacy applications
  - Allows leveraging of existing management infrastructure

IPoIB Configuration

- Requires assigning an IP address and a subnet mask to each HCA port (like any other network adapter)

- The first port on the first HCA in the host is called interface ib0, the second port is called ib1, and so on

- Configuration can be based on DHCP or on a static configuration
  - Modify /etc/sysconfig/network-scripts/ifcfg-ib0:

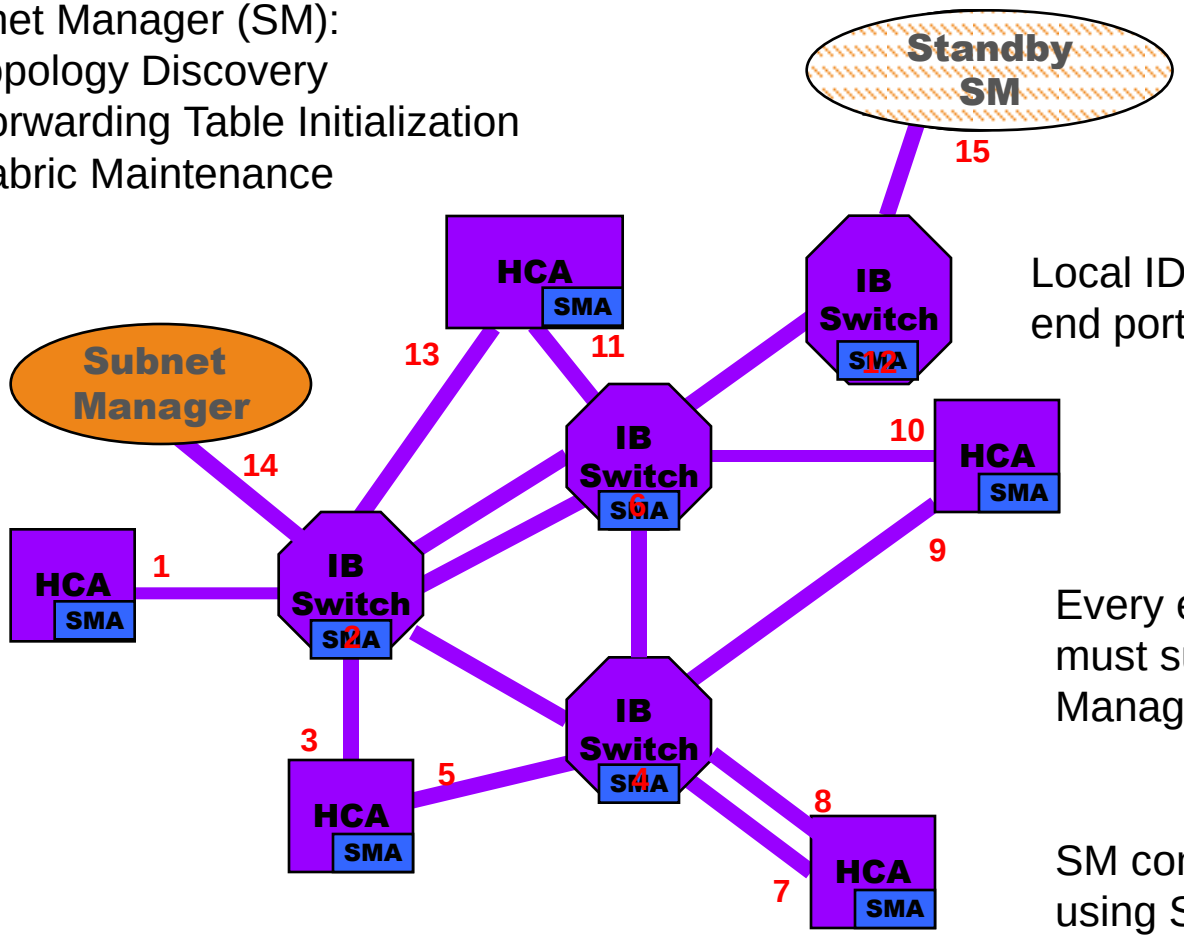
```
DEVICE=ib0
BOOTPROTO=static
IPADDR=10.10.0.1
NETMASK=255.255.255.0
NETWORK=10.10.0.0
BROADCAST=10.10.0.255
ONBOOT=yes
```

  - ifconfig ib0 10.10.0.1 up

Subnet Management Basics

Each Subnet must have a Subnet Manager (SM):
- Topology Discovery
- Forwarding Table Initialization
- Fabric Maintenance

Standby Subnet Managers Supported



Local IDs (LIDS) are used to identify end ports/nodes and route packets

Every entity (CA, SW, Router) must support a Subnet Management Agent (SMA)

SM communicates with SA using Subnet Management Packets (SMPs)

OpenSM in a nutshell

- The Subnet Manager (SM) is mandatory for setting up port ID, links and routes
- OpenSM is an Infiniband compliant subnet manger included with OFED
- Ability to run several instance of osm on the cluster in a Master/Slave(s) configuration for redundancy.
- Partitions P-key (similar to VLANs) support
- QoS support
- Enhanced routing algorithms:
  - Min-hop, up-down, fat-tree, LASH, DOR, Torus2QOS

Running opensm

- Command line
  - Default (no parameters)
    - Scans and initializes the IB fabric
  - opensm –h for usage flags
    - e.g to use up-down routing would using --routing_engine updn
  - Configuration file for advanced settings
    - -c creates config file
    - -F loads config file
  - Run is logged to two files
    - /var/log/messages – registers only major events
    - /var/log/opensm.log – detailed report
  - Can also start as a daemon using /etc/init.d/opensmd start

- SM detection
  - /etc/init.d/opensmd status
    - Shows opensm runtime status on a machine
  - sminfo
    - Shows master and standby subnets running on the cluster

Running Benchmarks

InfiniBand benchmarks

- Bandwidth and Latency performance tests
  - /usr/bin/ib_write_bw
  - /usr/bin/ib_write_lat
  - /usr/bin/ib_read_bw
  - /usr/bin/ib_read_lat
  - /usr/bin/ib_send_bw
  - /usr/bin/ib_send_lat
- Usage
  - Server: <test name> <options>
  - Client: <test name> <options> <server IP address>

Note: Same options must be passed to both server and client. Use –h for all options.

InfiniBand benchmarks

```
[root@lisbon001 ~]# ib_send_bw
```

```
[root@lisbon002 ~]# ib_send_bw lisbon001
---------------------------------------------------------------------
                    Send BW Test
Number of qps               : 1
Connection type             : RC
RX depth                    : 600
CQ Moderation               : 50
 Link type                  : IB
 Mtu                        : 2048
 Inline data is used up to 0 bytes message
 local address: LID 0x5c QPN 0x68004a PSN 0x821f36
 remote address: LID 0x5d QPN 0x64004a PSN 0xce92a9
---------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]
 65536      1000             2719.89                2719.47
---------------------------------------------------------------------
```

MPI benchmarks

- Prerequisites for Running MPI:
  - The mpirun_rsh launcher program requires automatic login (i.e., password-less) onto the remote machines.
  - Must also have an /etc/hosts file to specify the IP addresses of all machines that MPI jobs will run on.
  - Details on this procedure can be found in Mellanox OFED User's manual
- Basic format (mvapich):
  - mpirun_rsh –np *procs node1 node2 node3 BINARY*
  - See MPI User Manual for various tuning parameters for InfiniBand
    - E.g. Transport types at large scale,
- Other flags:

-show: show only

-paramfile: environment variables

-hostfile: list of host

-ENV=VAL (i.e. VIADEV_RENDEZVOUS_THRESHOLD=8000)

MPI benchmarks

```
[[root@lisbon001 ~]# mpirun_rsh -np 2 lisbon001 lisbon002
/usr/mpi/gcc/mvapich-1.2.0/tests/osu_benchmarks-3.1.1/osu_latency
# OSU MPI Latency Test v3.1.1
# Size          Latency (us)
0                1.37
1                1.36
2                1.37
4                1.37
8                1.38
16               1.38
32               1.45
64               1.55
128              2.43
256              2.56
512              2.86
1024             3.47
2048             4.75
4096             6.03
8192             8.76
16384           13.35
32768           18.46
65536           30.28
131072           52.84
262144           99.88
524288          191.46
1048576          375.02
2097152          748.70
4194304         1481.48
```

OFED InfiniBand Diagnostic Tools

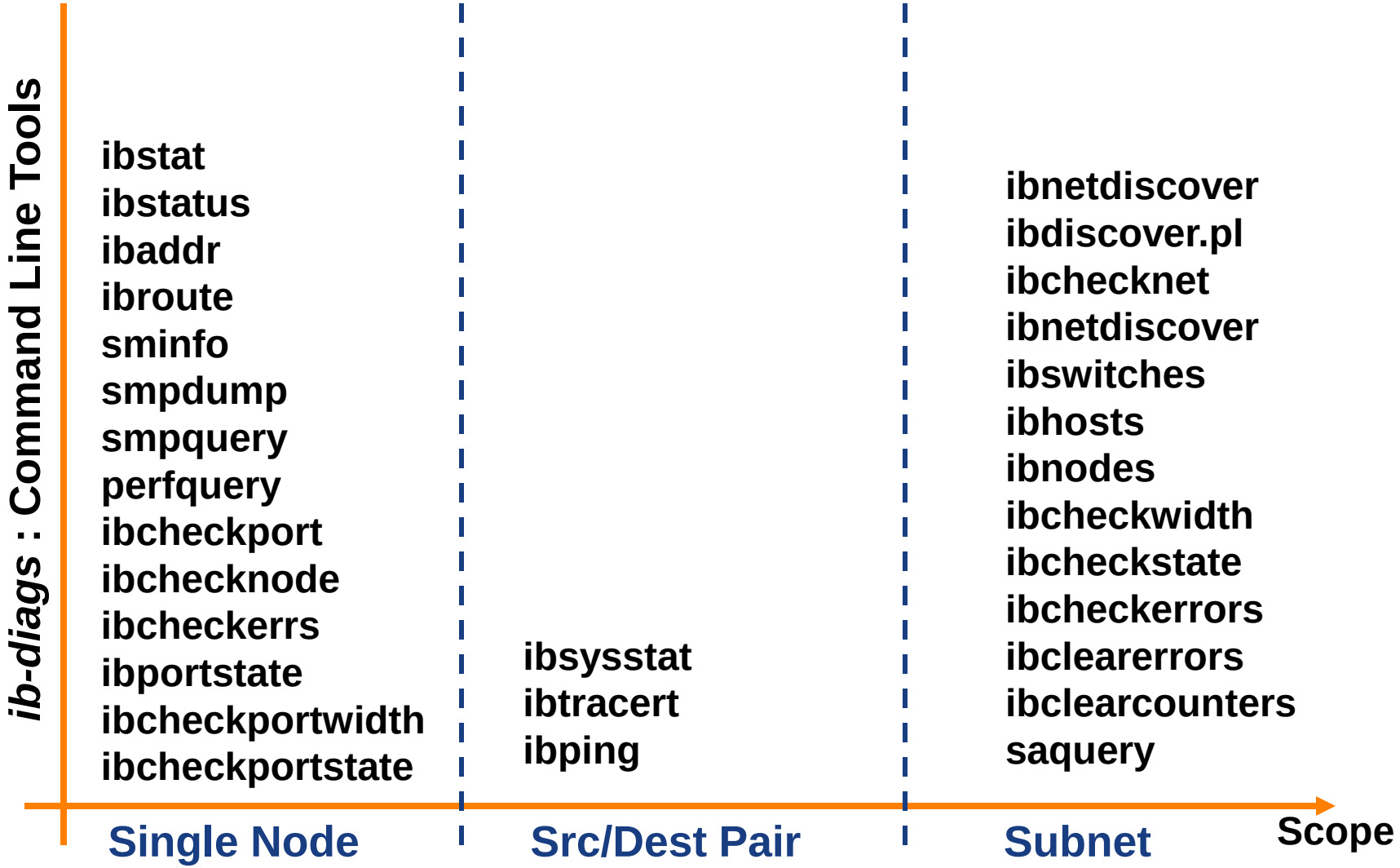
GUIDS – Globally Unique Device IDs

- All HCA and switches ship with GUIDs burned into non-volatile memory of the device
- The GUID is 64-bits (8-bytes) long
- Similar to MAC address in that they are globally unique and assigned by the manufacturer
- There are three types of GUIDS:
  - A single NODE GUID for every single switch or HCA device
  - A PORT GUID for every InfiniBand terminate port
    - E.g. Port 1 and Port 2 of an HCA, Port 0 (the management port) or a switch
  - A SYSTEM GUID - Allows a group of devices to be identified as a single unit
    - E.g. a large director class switch has a single SYSTEM GUID
- An example GUID: 0x0002c90200123456

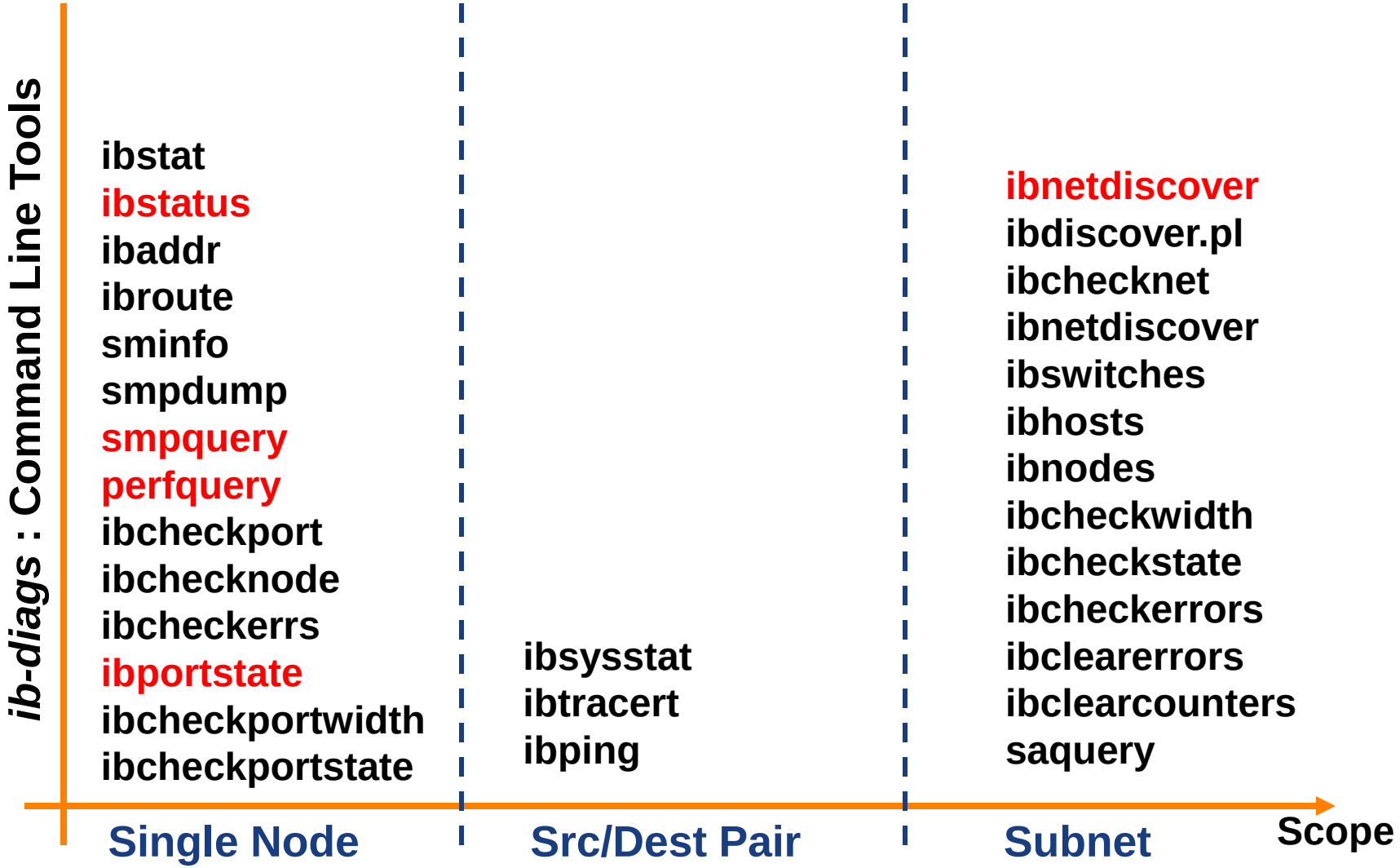
LIDS – Locally assigned IDs

- All HCA ports, and all Port 0 management ports on a switch are assigned LIDs
- These are dynamically assigned by the Subnet Manager, and can change from power cycle to power cycle (but not usually)
- The LID is 16-bits long: 64K LIDS, 48K unicast and 16K multicast
- LIDs are used to route InfiniBand packets through the subnet
- Example unicast LID: 0x1000 or 4096

The world of *ib-diags*



The world of *ib-diags*



The world of *ib-diags*

- Single Node Scope
  - ibstat - show host adapters status
  - ibstatus - similar to ibstat but implemented as a script
  - ibaddr - shows the lid range and default GID of the target (default is the local port)
  - ibroute - display unicast and multicast forwarding tables of switches
  - sminfo - query the SMInfo attribute on a node
  - smpdump - simple solicited SMP query tool. Output is hex dump
  - smpquery - formatted SMP query tool
  - perfquery - dump (and optionally clear) performance/error counters of the destination port
  - ibcheckport - perform some basic tests on the specified port
  - ibchecknode - perform some basic tests on the specified node
  - ibcheckerrs - check if the error counters of the port/node have passed some predefined thresholds
  - ibportstate - get the logical and physical port state of an IB port or enable/disable port
  - ibcheckportwidth - perform 1x port width check on specified port
  - ibcheckportstate - perform port state (and physical port state) check on specified port

- Node based tools can be run on any machine with OFED stack installed
  - man pages available for all utilities
  - -h option for online help

The world of *ib-diags*

Source/Destination Path Scope

- ibsysstat - obtain basic information for node (hostname, cpus, memory) which may be remote
- Ibtracert - display unicast or multicast route from source to destination
- ibping - ping/pong between IB nodes (currently using vendor MADs)

Subnet Scope

- saquery - issue some SA queries
- ibnetdiscover - scan topology
- ibchecknet - perform port/node/errors check on the subnet.
- ibnetdiscover - topology output
- ibswitches - scan the net or use existing net topology file and list all switches
- ibhosts - scan the net or use existing net topology file and list all hosts
- ibnodes - scan the net or use existing net topology file and list all nodes
- ibcheckwidth - perform port width check on the subnet. Used to find ports with 1x link width.
- ibclearerrors - clear all error counters on subnet
- ibclearcounters - clear all port counters on subnet
- ibcheckstate - perform port state (and physical port state) check on the subnet.
- ibcheckerrors - perform error check on subnet. Find ports above the indicated thresholds

ibstatus - HCA device information

- *ibstatus*
  - Displays basic information obtained from the local IB driver
  - Output includes: LID, SMLID, link width active, and port state

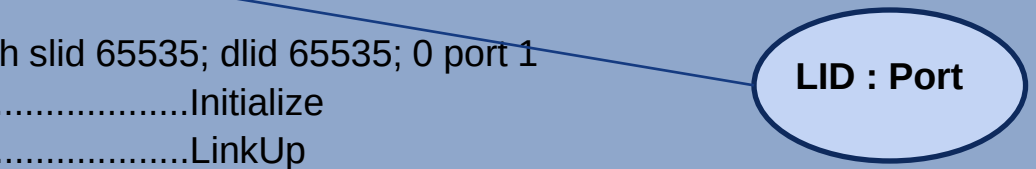
```
> ibstatus
Infiniband device 'mlx4_0' port 1 status:
        default gid:     fe80:0000:0000:0000:0000:0000:0007:3896
        base lid:        0x3
        sm lid:          0x3
        state:           4: ACTIVE
        phys state:      5: LinkUp
        rate:            20 Gb/sec (4X DDR)

Infiniband device 'mlx4_0' port 2 status:
        default gid:     fe80:0000:0000:0000:0000:0000:0007:3897
        base lid:        0x1
        sm lid:          0x3
        state:           4: ACTIVE
        phys state:      5: LinkUp
        rate:            20 Gb/sec (4X DDR)
```

Down = Physical Link is Down
Initialize = SM has not configured yet
Active = Ready to transfer data

ibportstate – port status/control

- *ibportstate*
  - Enables querying the logical (link) and physical port states of an InfiniBand port. It also allows adjusting the link speed that is enabled on any InfiniBand port
  - If the queried port is a switch port, then the command can also be used to:
    - Disable, enable or reset the port
    - Validate the port's link width and speed against the peer port

```
> ibportstate 56 3
PortInfo:
# Port info: DR path slid 65535; dlid 65535; 0 port 1
LinkState:.....Initialize
PhysLinkState:.....LinkUp
LinkWidthSupported:.....1X or 4X
LinkWidthEnabled:.....1X or 4X
LinkWidthActive:.....4X
LinkSpeedSupported:.....2.5 Gbps or 5.0 Gbps
LinkSpeedEnabled:.....2.5 Gbps or 5.0 Gbps
LinkSpeedActive:.....5.0 Gbps
```



perfquery – performance counter control

- *perfquery*
  - Queries InfiniBand ports' performance and error counters
  - It can also reset counters

```
perfquery
# Port counters: Lid 6 port 1
PortSelect:.....1
CounterSelect:.....0x1000
SymbolErrors:.....0
LinkRecovers:.....0
LinkDowned:.....0
RcvErrors:.....0
RcvRemotePhysErrors:.....0
RcvSwRelayErrors:.....0
XmtDiscards:.....0
XmtConstraintErrors:.....0
RcvConstraintErrors:.....0
LinkIntegrityErrors:.....0
ExcBufOverrunErrors:.....0
VL15Dropped:.....0
XmtData:.....55178210
RcvData:.....55174680
XmtPkts:.....766366
RcvPkts:.....766315
```

smpquery – attribute details

- *smpquery*
  - Reports relevant node, port, switch info

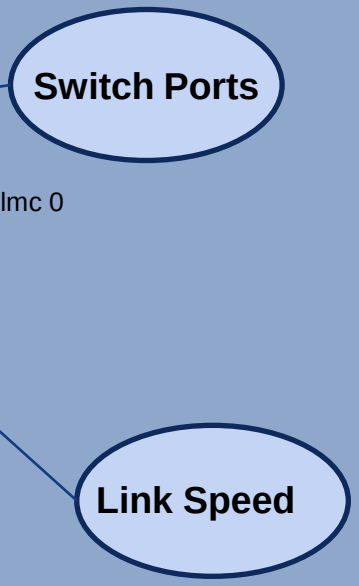
```
-> smpquery 5 6
Lid:.....0
SMLid:.....0
CapMask:.....0x0
DiagCode:.....0x0000
MkeyLeasePeriod:.....0
LocalPort:.....10
LinkWidthEnabled:.....1X or 4X
LinkWidthSupported:.....1X or 4X
LinkWidthActive:.....4X
LinkSpeedSupported:.....2.5 Gbps or 5.0 Gbps or 10.0 Gbps
LinkState:.....Active
PhysLinkState:.....LinkUp
LinkDownDefState:.....Polling
ProtectBits:.....0
LMC:.....0
LinkSpeedActive:.....10.0 Gbps
LinkSpeedEnabled:.....2.5 Gbps or 5.0 Gbps or 10.0 Gbps
NeighborMTU:.....2048
VLCap:.....VL0-7
…….continued
```

ibnetdiscover – cluster topology report

- Reports a complete topology of cluster
- Shows all interconnect connections reporting:
  - Port LIDs
  - Port GUIDs
  - Host names
  - Link Speed
- GUID to switch name file can be used for more readable topology

ibnetdiscover – cluster topology report

```
>  ibnetdiscover –node-name-map my_guid_map_file
     vendid=0x2c9
     devid=0xb924
     sysimgguid=0xb8cffff004207
     switchguid=0xb8cffff004207(b8cffff004207)
     Switch  24 SWITCH_1       # "MT47396 Infiniscale-III Mellanox Technologies" base port 0 lid 9 lmc 0
     [5]    "H-0002c90200230e54"[1](2c90200230e55)         # "mtilab55 HCA-1" lid 22 4xDDR
     [6]     "H-0002c902002312a8"[1](2c902002312a9)         # "mtilab47 HCA-1" lid 12 4xDDR
     [14]    "H-0002c90300000268"[2](2c9030000026a)         # "mtilab40 HCA-1" lid 20 4xDDR
     [18]    "H-0002c9020021ad78"[1](2c9020021ad79)         # "mtilab54 HCA-1" lid 21 4xDDR

     devid=0x6282
     sysimgguid=0x2c902002312ab
     caguid=0x2c902002312a8
     Ca      2 "H-0002c902002312a8"          # "mtilab47 HCA-1"
     [1](2c902002312a9)      "S-000b8cffff004207"[6]         # lid 12 lmc 0 "MT47396 Infiniscale-III Mellanox Technologies" lid 9 4xDDR

     vendid=0x2c9
     devid=0x6274
     sysimgguid=0x2c90200230e57
     caguid=0x2c90200230e54
     Ca      1 "H-0002c90200230e54"          # "mtilab55 HCA-1"
     [1](2c90200230e55)      "S-000b8cffff004207"[5]         # lid 22 lmc 0 "MT47396 Infiniscale-III Mellanox Technologies" lid 9 4xDDR
```



ibutils (ibdiagnet/path)
Integrated Cluster Utilities

Ibutils – in a nutshell

- *ibdiagnet*
  - Examine all the paths in the network
    - Look for cross paths issues
    - Network balancing
    - Covers all L2 issues on links
    - Latest MLNX_OFED stack includes ibdiagnet2, which includes performance optimizations and many advanced features

- *ibdiagpath*
  - Source to Destination path based analysis
    - Cover all L2 issues on the path
    - Include extensive link level analysis

ibdiagnet – functionality

- Topology
  - Info: dump topology in “topo”, “lst” and “ibnetdiscover” formats
  - Info: Optionally report on cable information (i.e. vendor, cable length, part number)
  - Error: duplicate GUIDs
  - Error: connectivity mismatch to reference topology
  - Warn: link speed/width change from reference topology
  - Error: optional report on any port below given speed/width
- SM
  - Info: all active SMs their status and priority
  - Error: missing or multiple masters
  - Error: Illegal LID: 0, duplicated, not meeting LMC
  - Error: invalid link parameters: OpVLs, MTU
  - Error: link width/speed not matching maximal supported
- Error Counters
  - Info: a full dump of all IB port counters of the entire subnet
  - Error: error counters over some limit (user controlled)
  - Error: error counters increasing during the run

ibdiagnet – functionality

- Routing
  - Info: histogram of hops from CA to CA
  - Info: histogram of number of CA to CA paths on every port
  - Info: multicast groups and their members (include sender only)
  - Error: no unicast route between every CA to every other CA
  - Error: (on request) no unicast route between every CA/SW to every other CA/SW
  - Error: credit loops found (optionally include multicast)
  - Error: multicast routing loops, disconnects, garbage
- Partitions
  - Info: All partitions ports and membership status
  - Error: Mismatching host partitions and attached switches ingress port tables
- IPoIB
  - Info: available broadcast domains and their parameters and member end-points
  - Warn: sub-optimal domain parameters (rate to small, rate not met by some nodes)
- Bit Error Check
  - Error: given some threshold and time between samples

ibdiagnet – functionality

- QoS
  - Info: Admissible SL's on the path (including the details where they block etc)
  - Info: PathRecord for every SL (optionally limit by given ServiceID, DSCP and SL)
  - Error: no common SL to be used
  - Error: no PathRecord for given ServiceID, DSCP and SL
- Cable Reports
  - Reports vendor information, part number, cable length, etc

HW failures – reporting faults using topology comparison

- Case 1: remove 2 cables
  - SL2-1 P10 to SL1-5 P14
  - SL2-6 P19 to SL1-10 P23
- Case 2: remove hosts
  - H-49, H-12
- Case 3: remove a switch, or a FRU within a switch system

Case 1:

ibdiagnet -t `pwd`/network.top

```
-I---------------------------------------------------
-I- Topology matching results
-I---------------------------------------------------
  Missing cable connecting:SL2-1/P10 to:SL1-5/P14
  Missing cable connecting:SL2-6/P19 to:SL1-10/P23
```

Case 2:

ibdiagnet -t `pwd`/network.top

```
-I---------------------------------------------------
-I- Topology matching results
-I---------------------------------------------------
  Missing System:H-12(MT23108)
     Should be connected by cable from port: P1(H-12/U1/P1)
     to:SL1-1/P12(SL1-1/U1/P12)

  Missing System:H-49(MT23108)
     Should be connected by cable from port: P1(H-49/U1/P1)
     to:SL1-5/P1(SL1-5/U1/P1)
```

Case 3:

ibdiagnet -t `pwd`/network.top

```
-I---------------------------------------------------
-I- Topology matching results
-I---------------------------------------------------
  Missing System Board:SL1-1/leaf3
```

Helpful Hint – use ibdiagnet to write out a topology file

- Writing out the topology
  - Use –wt network.topo to generate a reference topology of your cluster
  - Replace switch GUIDs with actual switch names for easier readability

Cable/connector faults and test

- Link Faults
  - Bad cables need to be found in cluster bring-up
  - Error counters provide on every IB port report these issues

- Reporting Link Faults across Network
  - Error: When any port counter change rate > than threshold
  - Report: Entire set of counters for each port on the subnet

```
ibdiagnet -t `pwd`/network.topo
-I---------------------------------------------------
-I- PM Counters Info
-I---------------------------------------------------
-W- "H-37/P1" lid=0x0087 guid=0x0002c900000000ee dev=23108
      Performance Monitor counter     : Value
      port_rcv_errors                 : 0x307 (Increase by 34 during ibdiagnet scan.)
-W- "SL1-2/P11" lid=0x0008 guid=0x0002c90000000207 dev=47396
      Performance Monitor counter     : Value
      port_rcv_errors                 : 0xd1 (Increase by 5 during ibdiagnet scan.)
-W- "SL1-2/P16" lid=0x0008 guid=0x0002c90000000207 dev=47396
      Performance Monitor counter     : Value
      port_rcv_errors                 : 0x6c (Increase by 4 during ibdiagnet scan.)
-W- "SL1-4/P1" lid=0x000c guid=0x0002c9000000020b dev=47396
      Performance Monitor counter     : Value
      port_xmit_discard               : 0x307 (Increase by 34 during ibdiagnet scan.)
```

Cable/connector faults and test

```
ibdiagpath -t `pwd`/network.topo –n H-3
-I-------------------------------------------------
-I- PM Counters Info
-I-------------------------------------------------
-W- "SL1-1/U1/P3" lid=0x0002 guid=0x0002c90000000201 dev=47396
      Performance Monitor counter      : Value
      port_rcv_errors                  : 0xcd4 (Increase by 7 during ibdiagpath scan.)
```

```
ibdiagpath -t `pwd`/network.topo –n H-23
-I-------------------------------------------------
-I- PM Counters Info
-I-------------------------------------------------
-W- "SL1-2/U1/P11" lid=0x0008 guid=0x0002c90000000207 dev=47396
      Performance Monitor counter      : Value
      port_rcv_errors                  : 0x603 (Increase by 8 during ibdiagpath scan.)
```

Advanced ibutil Topics

Subnet Manager faults

- The Subnet Manager (SM) is mandatory for setting up port ID, links and routes

- Subnet Manager Reporting
  - One and only one “master” SM
    - Error: When no master or more then one master
    - Report: All master and standby SM ports
  - SM is responsible for configuring links
    - Error: When Neighbor MTU is not correctly set by SM
    - Error: If operational VLs does not match the other sides of the link
  - Packet routes are configured by the SM
    - Error: When not all nodes are assigned an unique address (LID)
    - Error: If routes from every nodes to every other node are not set
    - Error: If multicast routes for each member of each group are not proper
    - Error: If “credit loops” are caused by the routing

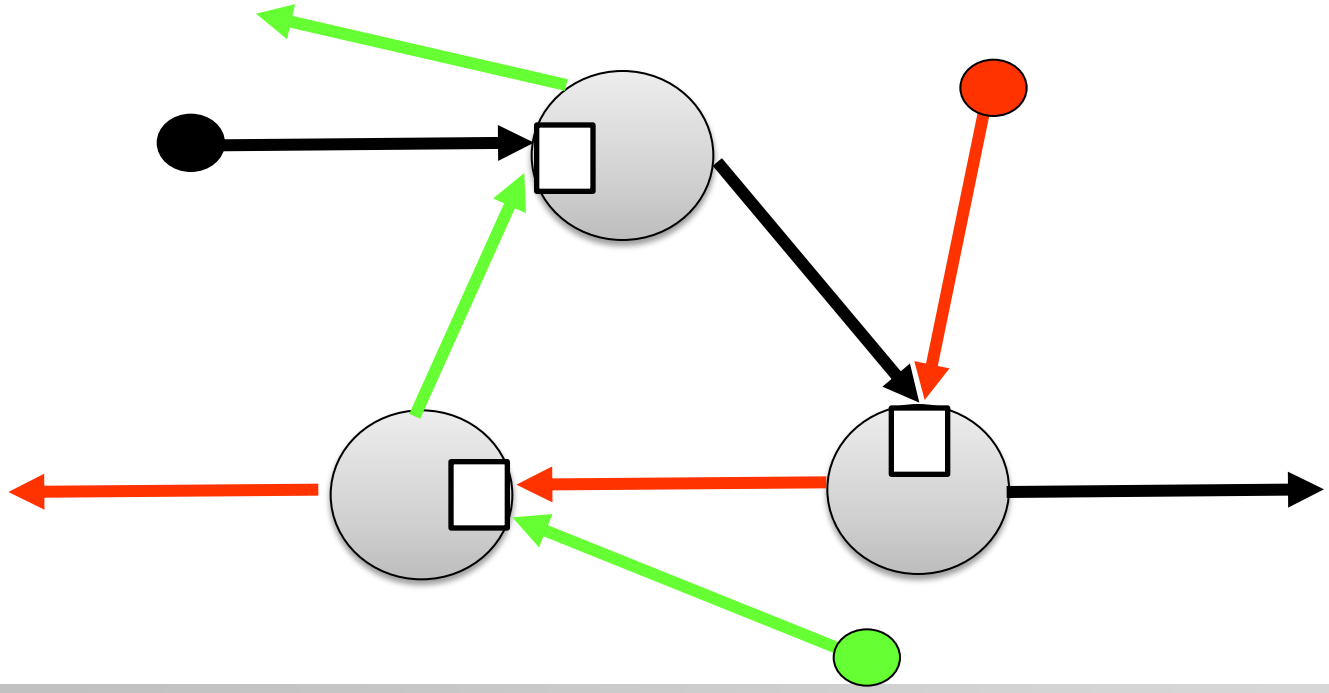
Broken route issues - *ibdiagnet* results

```
ibdiagnet -t `pwd`/network.topo -r
...
-I---------------------------------------------------
-I- Fabric qualities report
-I---------------------------------------------------
-I- Verifying all CA to CA paths ...
-E- Unassigned LFT for lid:70 Dead end at:H-122/U1
-E- Fail to find a path from:H-1/U1/1 to:H-24/U1/1
...
-E- Found 1380 missing paths out of:13340 paths
```

```
# Multicast disconnect and unneeded entries
-I- Scanning all multicast groups for loops and connectivity...
-I- Multicast Group:0xC000 has:6 switches and:8 FullMember CA ports
-W- Switch: S0002c90000000004/U1 has unconnected MFT entries for MLID:0xC000
-W- Switch: S0002c90000000005/U1 has unconnected MFT entries for MLID:0xC000
-E- Found 2 connection groups for MLID:0xC000
    Group:1 has 4 CAs: H-[9..12]/U1
    Group:1 has 1 SWs: S0002c90000000001/U1
    Group:2 has 4 CAs: H-[13..16]/U1
    Group:2 has 1 SWs: S0002c90000000003/U1
```

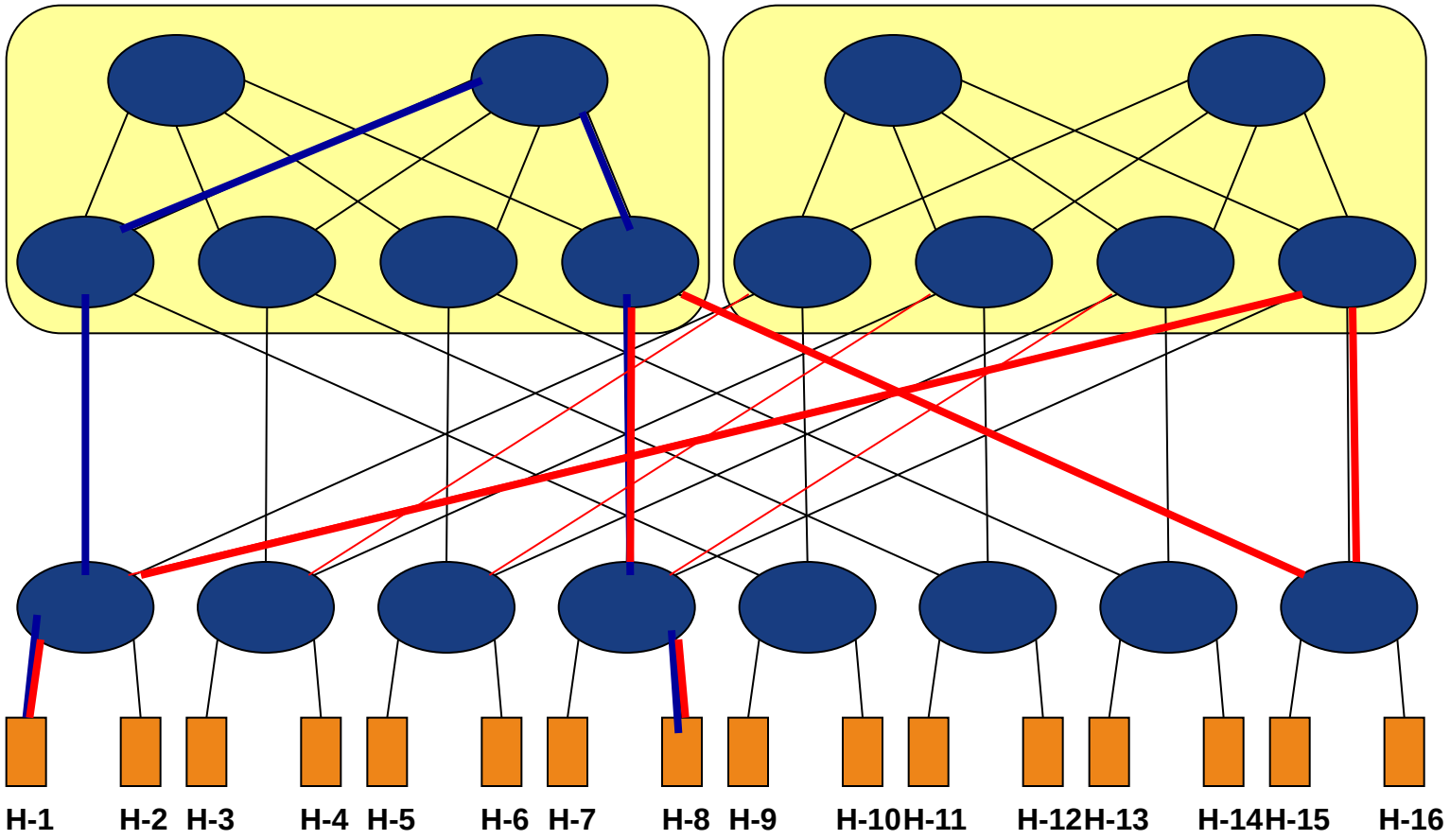
Credit Loops ? What are these?

- “loss-less fabric” = “link level flow control” = packet not sent of there is no buffer for it
- If traffic to DST-1 waits on traffic for DST-2 which in turn depends on traffic to DST-3 which depends on DST-1 we have a dependency loop and the fabric deadlocks



SM/SA Test Cases

- Credit Loops in real world



Credit Loop issues - *ibdiagnet* results

```
ibdiagnet -t `pwd`/network.topo -r
...
-I---------------------------------------------------
-I- Checking credit loops
-I---------------------------------------------------
-I-
-I- Analyzing Fabric for Credit Loops 1 SLs, 1 VLs used.
    Found credit loop on: SW_L2_1/P3 VL: 0
  - BT credit loop through: SW_L1_8/P2 VL: 0
  - BT credit loop through: SW_L2_8/P3 VL: 0
  - BT credit loop through: SW_L1_4/P2 VL: 0
  - BT credit loop through: SW_L2_4/P3 VL: 0
  - BT credit loop through: SW_L1_7/P2 VL: 0
  - BT credit loop through: SW_L2_7/P3 VL: 0
  - BT credit loop through: SW_L1_3/P2 VL: 0
  - BT credit loop through: SW_L2_3/P3 VL: 0
  - BT credit loop through: SW_L1_6/P2 VL: 0
  - BT credit loop through: SW_L2_6/P3 VL: 0
  - BT credit loop through: SW_L1_2/P2 VL: 0
  - BT credit loop through: SW_L2_2/P3 VL: 0
  - BT credit loop through: SW_L1_5/P2 VL: 0
  - BT credit loop through: SW_L2_5/P3 VL: 0
  - BT credit loop through: SW_L1_1/P2 VL: 0
  - BT credit loop through: SW_L2_1/P3 VL: 0
  - BT credit loop through: H-1/P1 VL: 0
-E- credit loops in routing

-E- Total Credit Loop Check Errors:1
```

Partition Configuration Issues

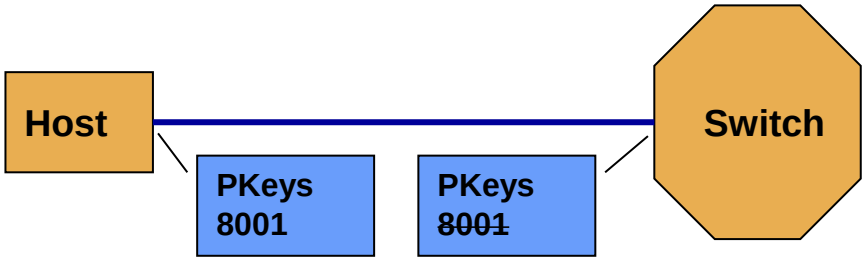
- Partitions are similar to VLAN IDs, but enforced on hosts and switch ports

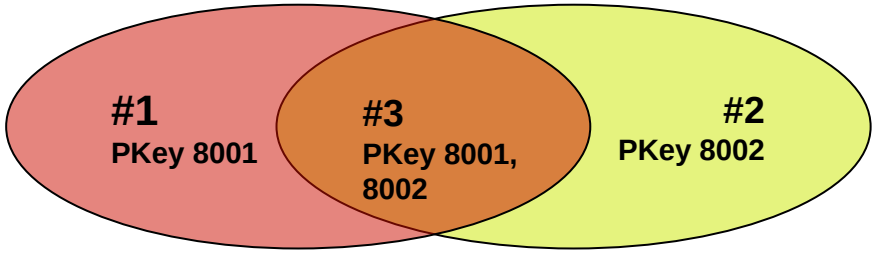
- Network
  - Warning: Partition enforcement by leaf switches mismatch hosts
  - Report: Node groups – which nodes can communicate

- Path
  - Error: No common partition for the path
  - Error: Mismatch between leaf switch and host partitions
  - Report: Which partitions can be used for the path
  - Verbose: On each port (if enforcing) show list of PKeys

Partitions Test Case

- Two partitions with some common nodes.






```
-I---------------------------------------------------
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I---------------------------------------------------
-W- Missing PKey:0x8001 on remote switch of node:"H-95/P1" lid=0x0089
    guid=0x0002c900000001ee dev=23108
-I-    PKey:0x0001 Hosts:87 full:87 limited:0
-I-    PKey:0x0002 Hosts:84 full:84 limited:0
-I-    PKey:0x7fff Hosts:128 full:1 limited:127
-I---------------------------------------------------
```

Recap - Verifying InfiniBand interconnect

- ibdiagnet is a simple to use diagnostic tools to monitor the quality of interconnect
- Once cluster is built run ibdiagnet to quickly scan and find initial problems
- Stress the cluster. MPI communication benchmarks are best for this
  - e.g. IMB
- Use an iterative process of stressing the cluster and then scanning for errors until all fabric problems are discovered and fixed
- Take snapshot of topology and use this to run periodic scans of the fabric
- Use point tools for problem isolation
  - e.g ibdiagpath, ibportstate, perfquery, etc

Questions?

Extra Info

ibdiagnet run – a good case

ibdiagnet -ls 10 –lw 4x

```
-I- Parsing Subnet file:/tmp/ibmgtsim.7021/ibdiagnet.lst
-I- Defined 145/145 systems/nodes
-I---------------------------------------------------
-I- Bad Guids/LIDs Info
-I---------------------------------------------------
-I- skip option set. no report will be issued
-I---------------------------------------------------
-I- Links With Logical State = INIT
-I---------------------------------------------------
-I- No bad Links (with logical state = INIT) were found
-I---------------------------------------------------
-I- General Device Info
-I---------------------------------------------------
-I- PM Counters Info
-I---------------------------------------------------
-I- No illegal PM counters values were found
-I---------------------------------------------------
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I---------------------------------------------------
-I-    PKey:0x0001 Hosts:128 full:128 limited:0
-I-    PKey:0x0002 Hosts:128 full:128 limited:0
-I-    PKey:0x0003 Hosts:128 full:128 limited:0
-I-    PKey:0x7fff Hosts:128 full:1 limited:127
-I---------------------------------------------------
-I- IPoIB Subnets Check
-I---------------------------------------------------
-I- Subnet: IPv4 PKey:0x0001 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0002 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0003 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I---------------------------------------------------
```

ibdiagnet run – a good case

```
-I---------------------------------------------------
-I- Bad Links Info
-I- No bad link were found
-I---------------------------------------------------
-I- Summary Fabric SM-state-priority
-I---------------------------------------------------
  SM - master
    The Local Device : H-1/P1 lid=0x0001 guid=0x0002c90000000002 dev=23108
    priority:0
-I---------------------------------------------------
-I- Fabric qualities report
-I---------------------------------------------------
-I- Parsing FDBs file:/tmp/ibmgtsim.7021/ibdiagnet.fdbs
-I- Defined 2465 fdb entries for:17 switches
-I- Parsing Multicast FDBs file:/tmp/ibmgtsim.7021/ibdiagnet.mcfdbs
-I- Defined 450 Multicast Fdb entries for:17 switches
-I-
-I- Verifying all CA to CA paths ...
    ---------------------- CA to CA : LFT ROUTE HOP HISTOGRAM -----------------
    The number of CA pairs that are in each number of hops distance.
    This data is based on the result of the routing algorithm.

    HOPS NUM-CA-CA-PAIRS
  2   1364
  4   14892
  ------------------------------------------------------------------------------
```

ibdiagnet run – a good case

```
---------- LFT CA to CA : SWITCH OUT PORT - NUM DLIDS HISTOGRAM -----------
 Number of actual Destination LIDs going through each switch out port
 considering
 all the CA to CA paths. Ports driving CAs are ignored (as they must
 have = Nca - 1). If the fabric is routed correctly the histogram
 should be narrow for all ports on same level of the tree.
 A detailed report is provided in /tmp/ibdmchk.sw_out_port_num_dlids.

 NUM-DLIDS NUM-SWITCH-PORTS
      1   20
      2   84
      3   21
      4   2
      5   1
      9   28
     10   72
     11   28
     ------------------------------------------------------------------------
-I- Scanned:16256 CA to CA paths
    ------------------------------------------------------------------------
-I- Scanning all multicast groups for loops and connectivity...
-I- Multicast Group:0xC000 has:12 switches and:128 HCAs
-I- Multicast Group:0xC001 has:12 switches and:128 HCAs
-I- Multicast Group:0xC002 has:12 switches and:128 HCAs
-I-------------------------------------------------
-I- Checking credit loops
-I-------------------------------------------------
-I- Analyzing Fabric for Credit Loops 1 SLs, 1 VLs used.
-I- no credit loops found
```

ibdiagnet run – a good case

```
-I---------------------------------------------------
-I- mgid-mlid-HCAs table
-I---------------------------------------------------
mgid                                    | mlid   | PKey   | QKey       | MTU   | rate     | HCAs
0xff12401b80010000:0x00000000ffffffff | 0xc000 | 0x8001 | 0x00000b1b | =2048 | =20Gbps  | 128
0xff12401b80020000:0x00000000ffffffff | 0xc001 | 0x8002 | 0x00000b1b | =2048 | =20Gbps  | 128
0xff12401b80030000:0x00000000ffffffff | 0xc002 | 0x8003 | 0x00000b1b | =2048 | =20Gbps  | 128
----------------------------------------------------------------
-I- Stages Status Report:
    STAGE                                Errors Warnings
    Bad GUIDs/LIDs Check                 0      0
    Link State Active Check              0      0
    General Devices Info Report          0      0
    Performance Counters Report          0      0
    Partitions Check                     0      0
    IPoIB Subnets Check                  0      0
    Subnet Manager Check                 0      0
    Fabric Qualities Report              0      0
    Credit Loops Check                   0      0
    Multicast Groups Report              0      0

Please see /tmp/ibmgtsim.7021/ibdiagnet.log for complete log
----------------------------------------------------------------
```

Ibdiagpath run – a good case

ibdiagpath -t network.topo -l 128

```
-I- Parsing topology definition:/local/ez/OSM_REGRESSION/SRC/ibutils/ibdiag/demo/network.topo
-I- Defined 145/145 systems/nodes
-I-------------------------------------------------
-I- Traversing the path from local to destination
-I-------------------------------------------------
-I- From: "H-1/U1/P1"    lid=0x0001 guid=0x0002c90000000002 dev=23108
-I- To:   "SL1-1/U1/P1"  lid=0x0002 guid=0x0002c90000000201 dev=47396

-I- From: "SL1-1/U1/P21" lid=0x0002 guid=0x0002c90000000201 dev=47396
-I- To:   "SL2-5/U1/P1"  lid=0x0020 guid=0x0002c9000000021f dev=47396

-I- From: "SL2-5/U1/P5"  lid=0x0020 guid=0x0002c9000000021f dev=47396
-I- To:   "SL1-3/U1/P21" lid=0x0009 guid=0x0002c90000000209 dev=47396

-I- From: "SL1-3/U1/P9"  lid=0x0009 guid=0x0002c90000000209 dev=47396
-I- To:   "H-33/U1/P1"   lid=0x0080 guid=0x0002c900000000de dev=23108
-I-------------------------------------------------
-I- PM Counters Info
-I-------------------------------------------------
-I- No illegal PM counters values were found
-I-------------------------------------------------
-I- Path Partitions Report
-I-------------------------------------------------
-I- Source "H-1/U1/P1" lid=0x0001 guid=0x0002c90000000002 dev=23108 Port 1
    PKeys:0xffff 0x8001 0x8002 0x8003
-I- Destination "H-33/U1" lid=0x0080 guid=0x0002c900000000de dev=23108
    PKeys:0x7fff 0x8001 0x8002 0x8003
-I- Path shared PKeys: 0x8001 0xffff 0x8002 0x8003
```

Ibdiagpath run – a good case

```
-I---------------------------------------------------
-I- IPoIB Path Check
-I---------------------------------------------------
-I- Subnet: IPv4 PKey:0x0001 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0002 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I- Subnet: IPv4 PKey:0x0003 QKey:0x00000b1b MTU:2048Byte rate:20Gbps SL:0x00
-I---------------------------------------------------
-I- QoS on Path Check
-I---------------------------------------------------
-I- The following SLs can be used:0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
-----------------------------------------------------------------
-I- Stages Status Report:
    STAGE                                   Errors Warnings
    LFT Traversal: local to destination     0      0
    Performance Counters Report             0      0
    Path Partitions Check                   0      0
    Path IPoIB Check                        0      0
    QoS on Path Check                       0      0

Please see /tmp/ibdiagpath.log for complete log
```